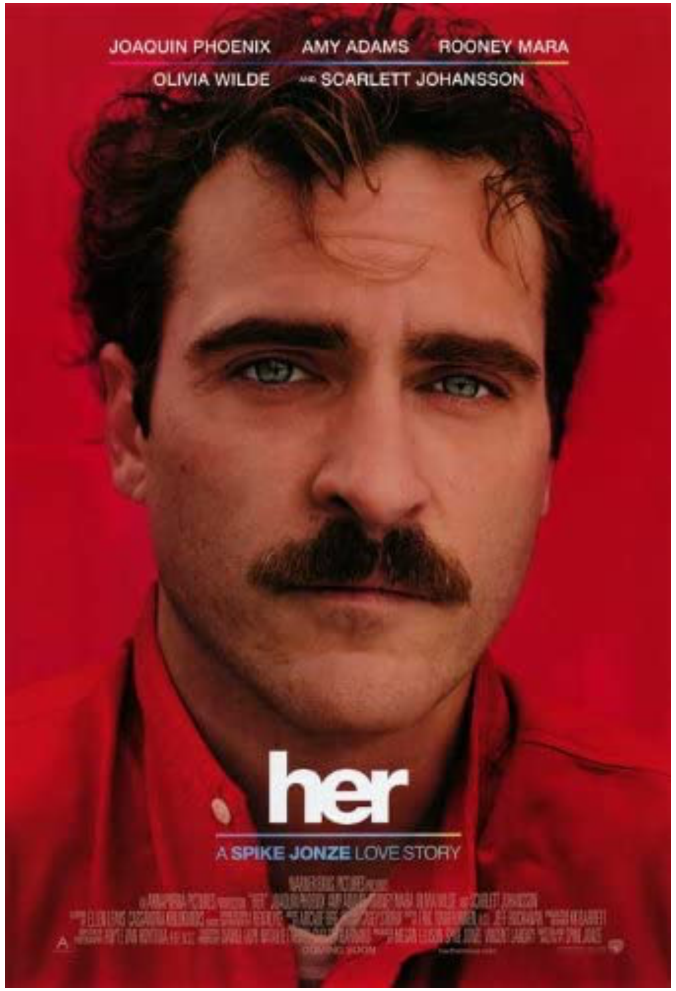
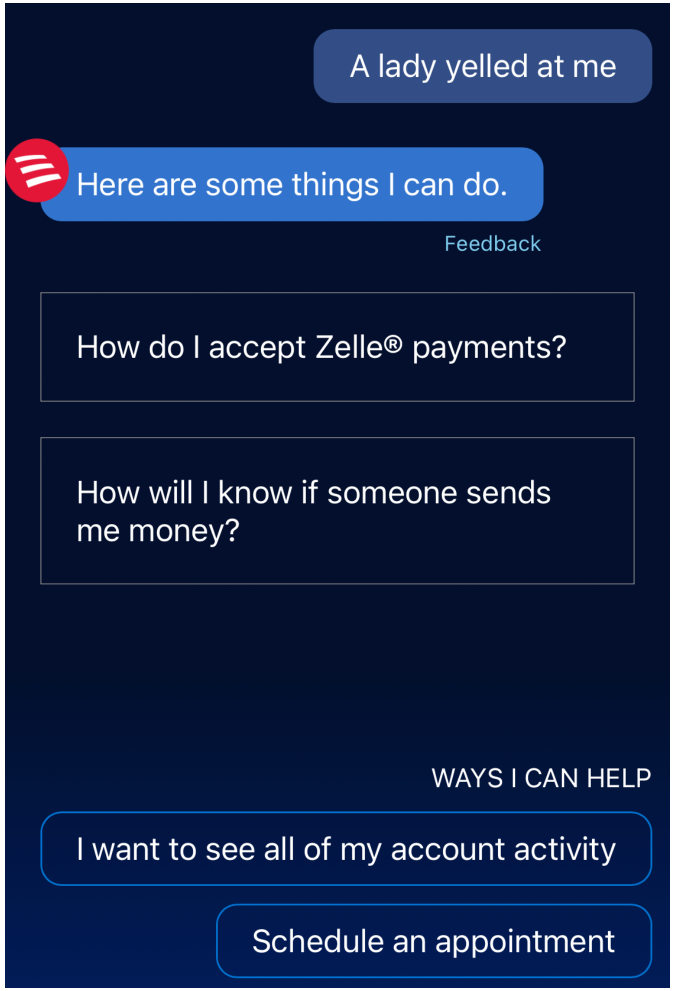
Jokes or Not

Cloudy Liu

Ideal vs Reality


JOAQUIN PHOENIX AMY ADAMS ROONEY MARA
OLIVIA WILDE AND SCARLETT JOHANSSON
her
A SPIKE JONZE LOVE STORY



A lady yelled at me

Here are some things I can do.

Feedback

How do I accept Zelle® payments?

How will I know if someone sends me money?

WAYS I CAN HELP

I want to see all of my account activity

Schedule an appointment

Problem Statement

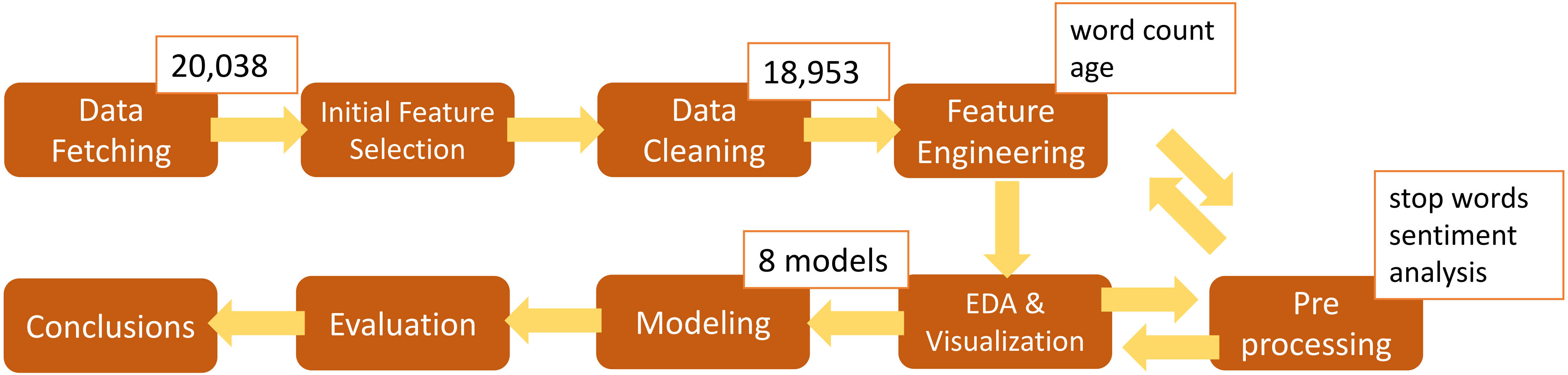
- To build a classifier that is able to tell an intended joke from the sharing of a personal story
- To serve as a first step to give virtual assistants some personal touch during the communication with customers
- So that to establish emotional bonds down the way
- Audience: product managers at an online retail company

Data Selection

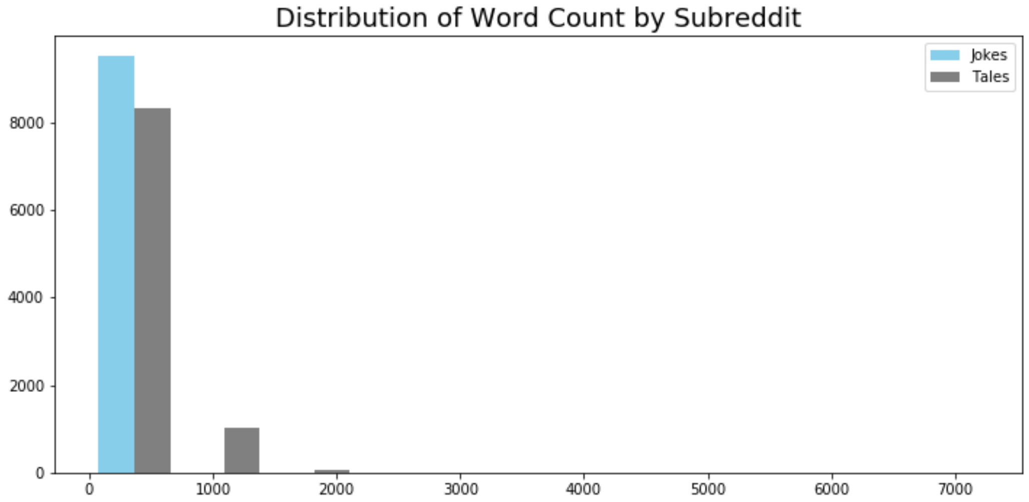
Subreddits: Jokes(19.4m) vs TalesfromRetail (604k)

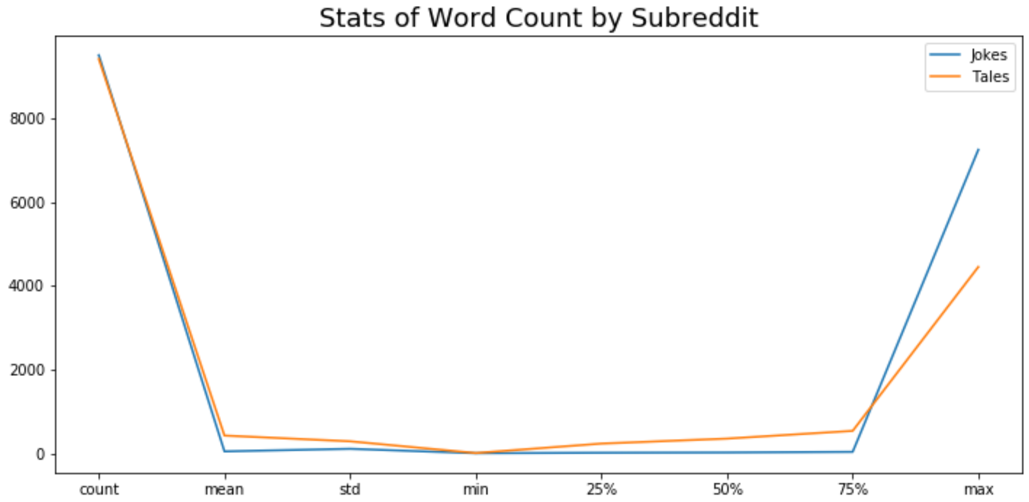
- Both could be very funny
- However, one with an intention to entertain (punch line), the other was to share (loose structure)

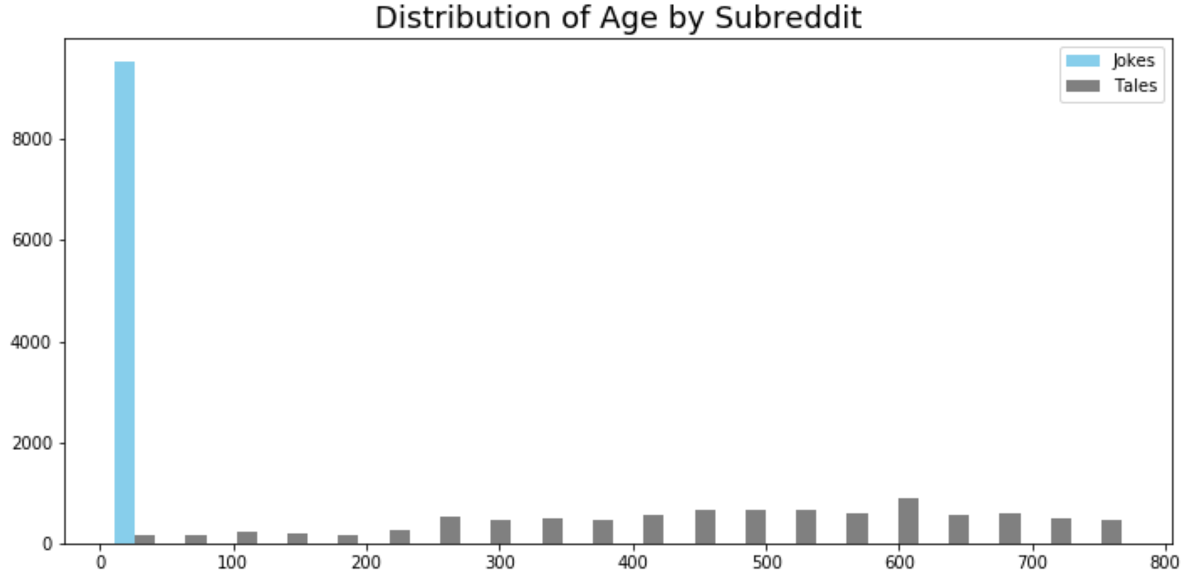
Process


Data Fetching
20,038
Initial Feature Selection
Data Cleaning
18,953
Feature Engineering
word count
age
Pre processing
stop words
sentiment analysis
EDA & Visualization
Modeling
8 models
Evaluation
Conclusions

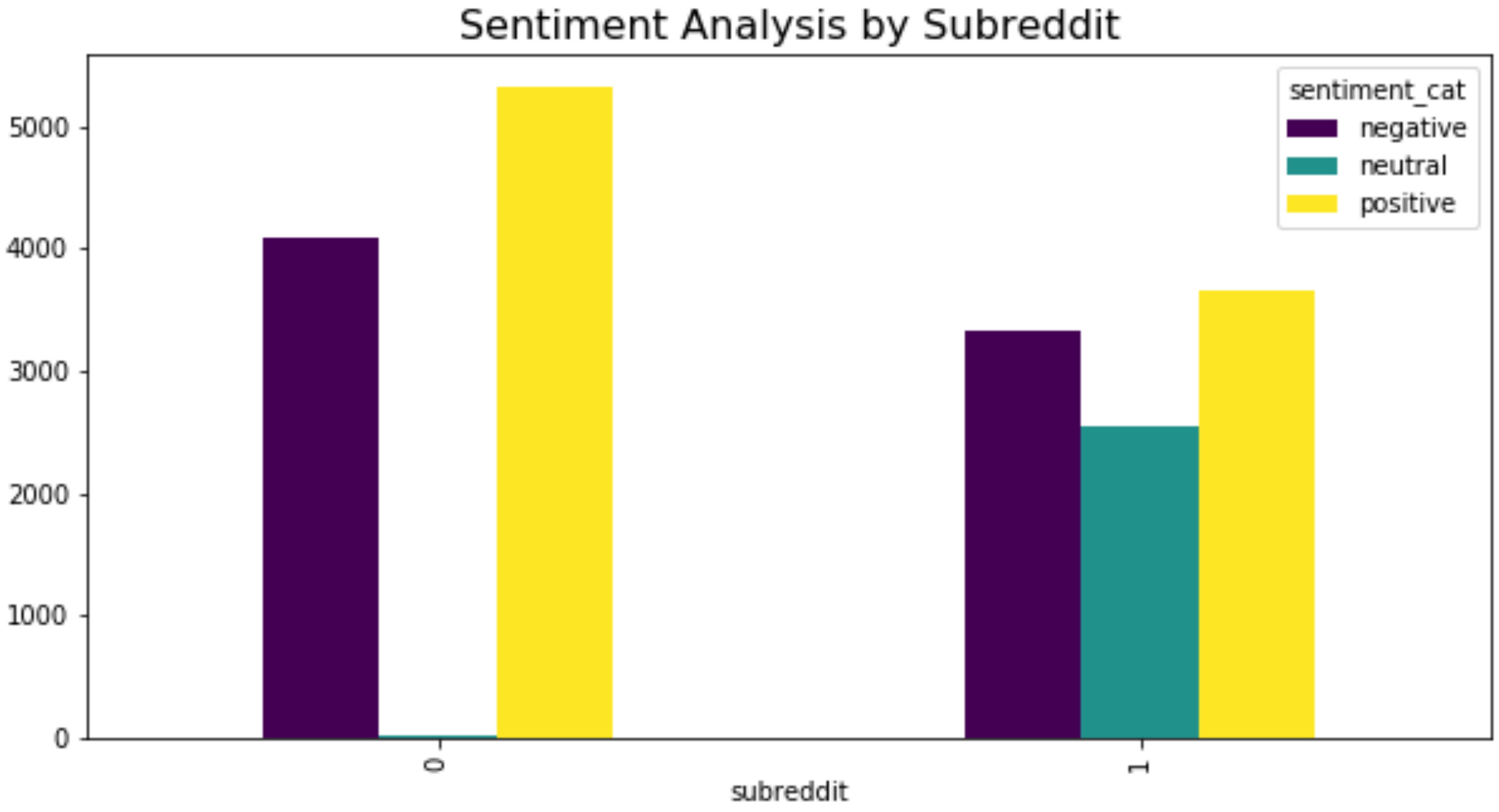
Exploratory Data Analysis


Distribution of Word Count by Subreddit


Stats of Word Count by Subreddit


Distribution of Age by Subreddit

Sentiment Analysis



Model Goal Setting

What metrics do I care the most?

- Accuracy

You want to say the right things to the right mood

- Specificity

You don’t want to make fun of people’s bad experience (rant/complaint…)

Other factors?

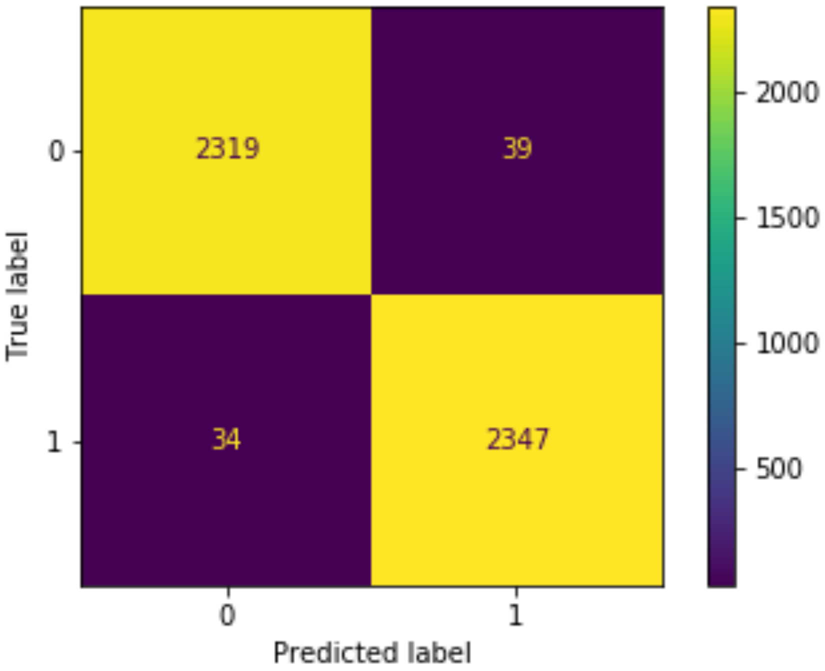
- Feature Numbers
- Computational Costs

Model Comparison

%	Baseline	Logistic Regression	K-Nearest Neighbors	Naïve Bayes	Random Forest	Extra Trees
	50.2	10k, 4k		4k		
CV		99		92.5		
TV		98.5	93.6	93	98.3	97.8
		2k	4k	4k	2k	2k

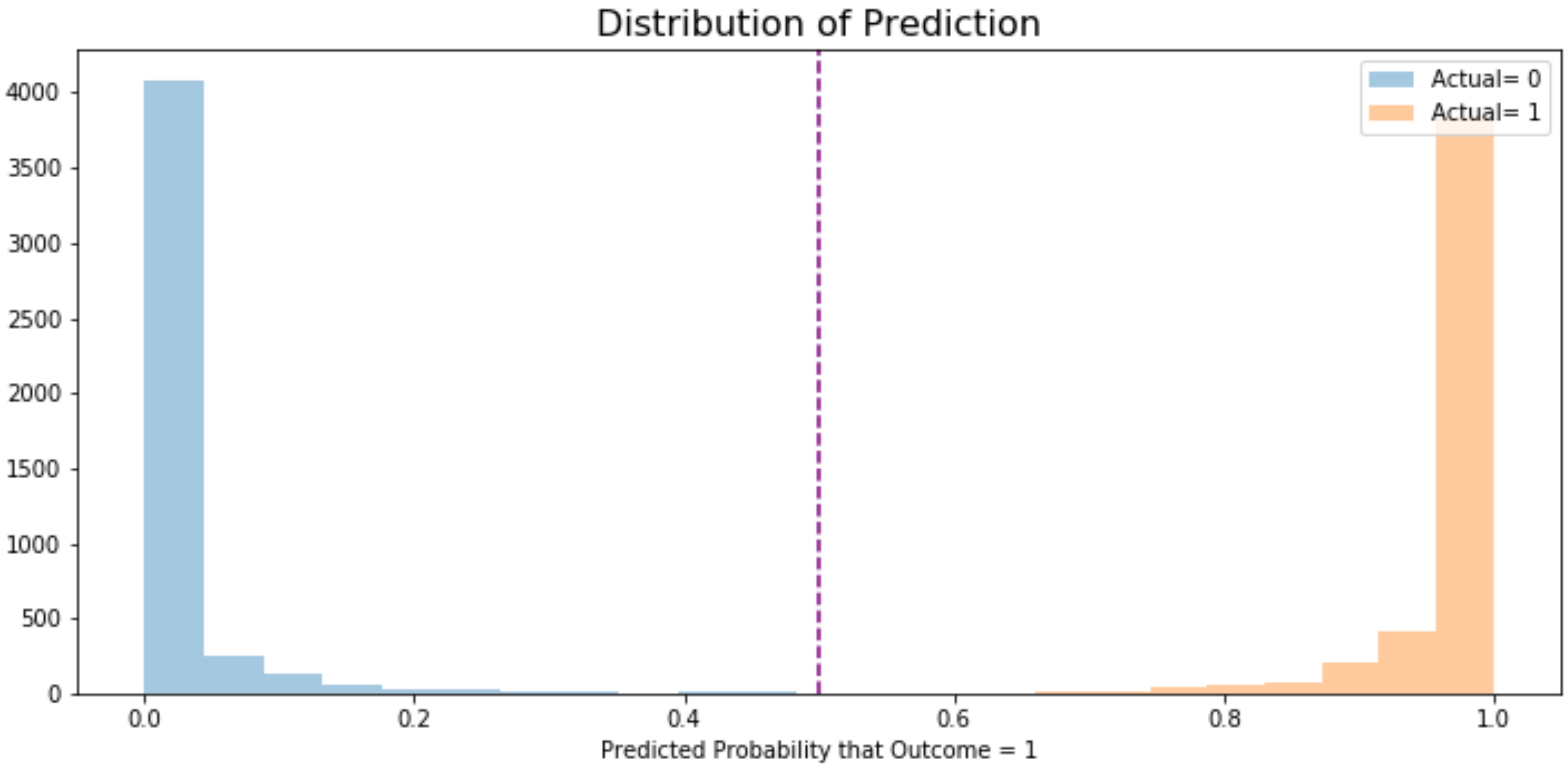
Best Model Selection

Logistic Regression with TfidfVectorizer



- High Accuracy: 98.5%
- High Specificity: 98.3%
- Fewer Features : 2,000
- Faster Run Time: 16min for 120 fits
- Interpretability

Evaluation


Distribution of Prediction
Actual= 0
Actual= 1
4000
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
Predicted Probability that Outcome = 1

Jokes or Tales?

Jokes

- I was told to leave the grocery store. They said I was taking way too long at the self-checkout. It's not my fault........I didn't put the mirror there.

Tales

- Easy Like Sunday Morning Customer: “Will you call me a cab?” Me: “Sure! You’re a cab! Who’s next, please?”

Tales

- I run a small online store, we sell custom hair extensions. Client orders, to put it in her words "The blondest blonde you have to match my hair color that is really blonde you know like blonde blonde” So we dye the hair to lightest ash blonde, manufacture to specification and courier it out. Client sends an email "I'm soooo upset, I ordered blonde this is NOT blonde”. I reply that I personally did the color work, I know it was as blonde as possible, please send a picture of the item she received to make sure the right item was dispatched. She sends a picture of exactly the correct, lightest ash blonde item we make. Now I'm confused, need to understand this further and phone her. She answers and packs out laughing "What's wrong?" I ask, "Nothing, all good" she says. "Please I need to understand what was wrong with the color?” Then she answered with the blondest line I've ever heard. "I had my sunglasses on”

Misclassification

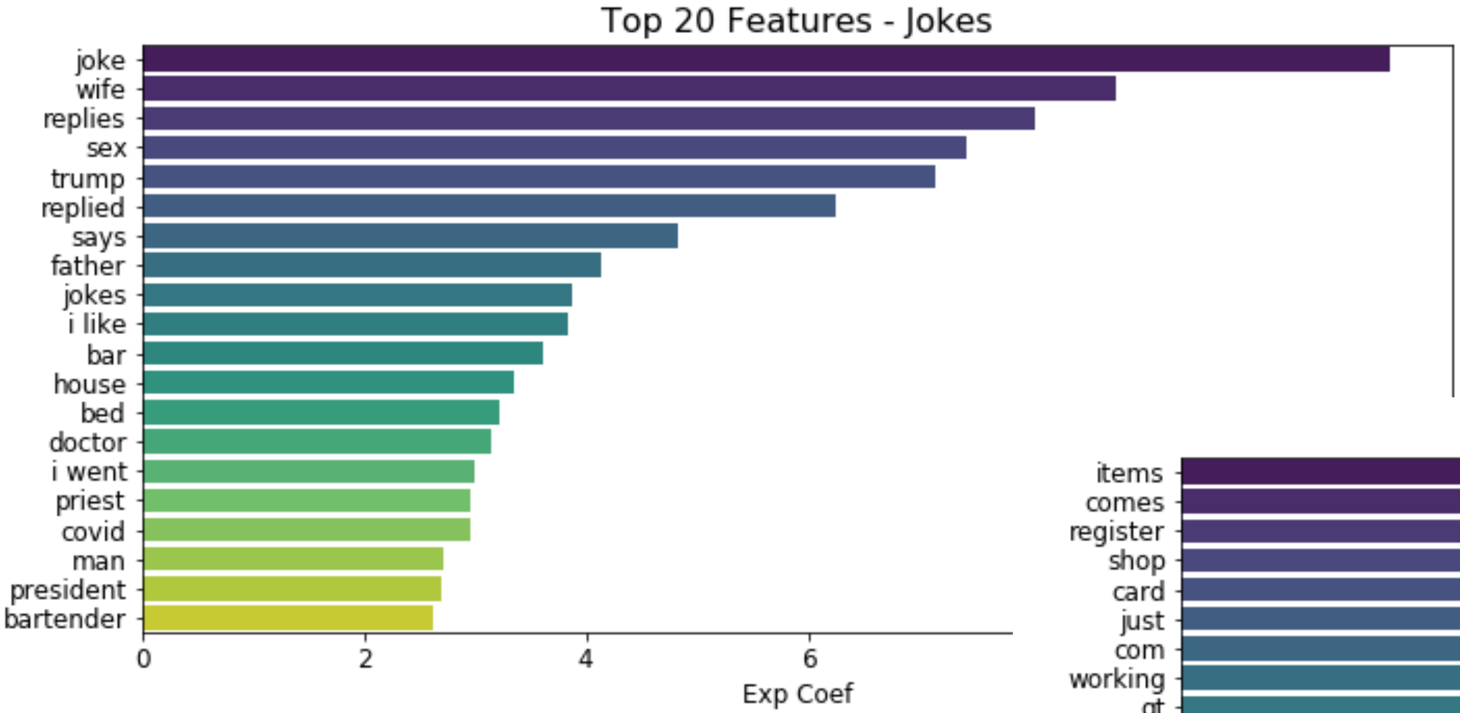
Tales as Jokes(39)

- Positive: 23
- Neutral: 0
- Negative: 16
- Avg word count: 125
- Max word count: 309

Jokes as Tales (34)

- Positive: 16
- Neutral: 10
- Negative: 8
- Avg word count: 178
- Max word count: 1241

Significant Features

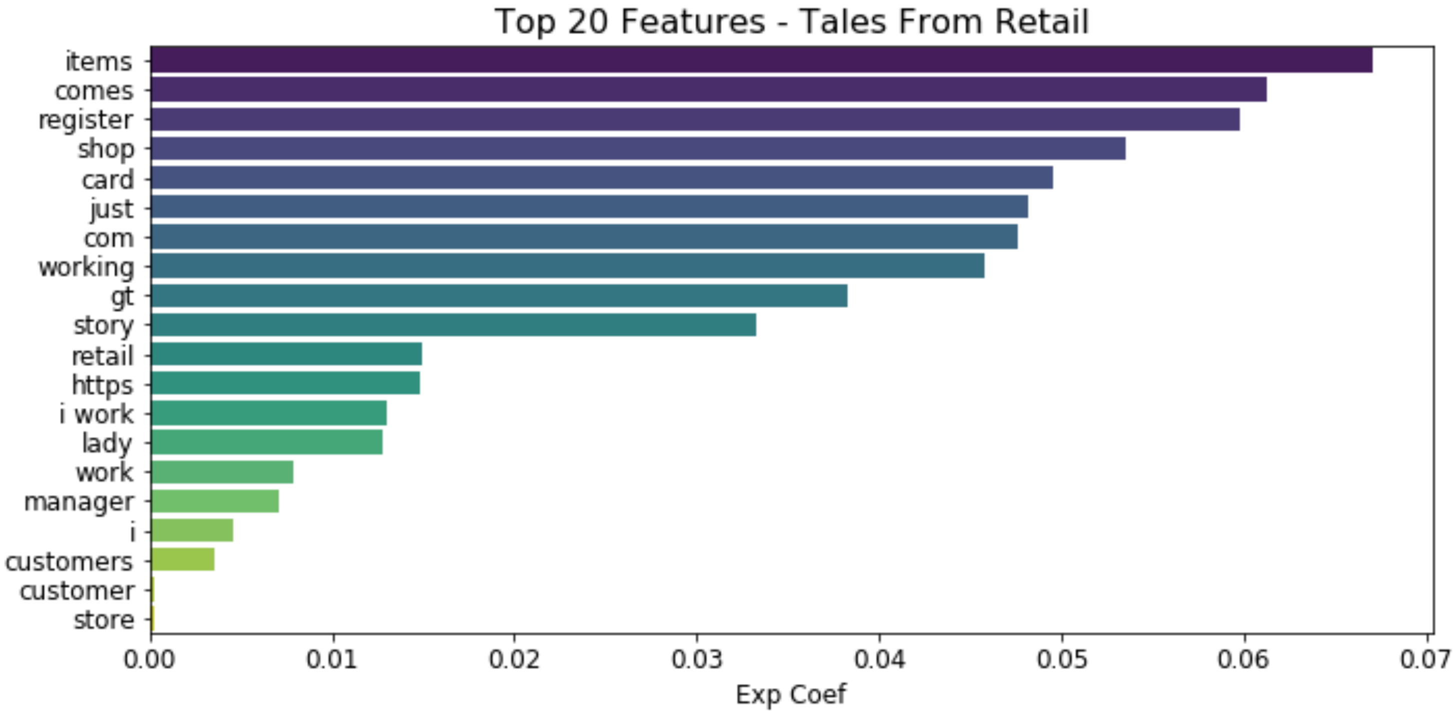


Role/Profession

Sexual

In-trend

Retail



Conclusions

- The two subreddits are very likely distinctive enough;
- Logistic Regression with TfidVectorizer provides the best model with high accuracy and high specificity. However it also tends to classify based on the length of the texts;
- Sentiment analysis could be used as a tool to improve classification;
- Key identifiers for Jokes: professions (”priest”, “doctor”, “bartender”, “president”), sexual(”sex”, “bed”, “wife”), trend words(“trump”, “covid”);
- Key identifiers for TalesfromRetail: retail terms (“items”, “registers”, “customer(s)”, “manager”, “I work”).